PEER-REVIEWED JOURNAL ON THE INTERNET

PEER-REVIEWED JOURNAL ON THE INTERNET

In the realm of Big Data ...

by Antony Bryant and Uzma Raja

Abstract

In 2008, Chris Anderson (2008), at that time the Editor–in–Chief of *Wired*, proposed that in the age of the petabyte, there was no longer any need for the scientific method, nor for models or theories. Although it might be contended that this was more provocation and journalistic hubris than formal or substantiated claim, the issue was taken up and has gathered momentum ever since. Indeed within a year or so of Anderson's article, and a series of rejoinders published on the Edge Web site, 'The Age of Big Data' was being heralded, and the measure had increased from petabytes to exabytes, zettabytes, and yottabytes. Diebold (2012) usefully distinguishes between Big Data 'the phenomenon', 'the term', and 'the discipline'; arguing that the phenomenon 'continues unabated', the term is 'firmly entrenched', and the discipline is 'emerging'. In what follows we focus initially on the term and the phenomenon, but our main objective is to argue that it is critical that there is general understanding of the emerging discipline. In particular we aim to justify the assertion that in the age of Big Data the ability to be able to develop abstractions and concepts is at least as important as it was previously; perhaps even more so. Moreover that these skills and techniques need to be understood and available to all of us in an era where we are all analysts and researchers at least to the extent of our use of the internet and its potential for affording search and investigation of online resources. We seek to offer some critical insights into these activities — modeling, conceptualizing, and theorizing — by comparing and contrasting Knowledge Discovery from Data (KDD) with the Grounded Theory Method (GTM). The former a technical orientation, that although predating Big Data, lies at the heart of the emerging tools and techniques. The latter a widely used approach to qualitative research aimed at developing conceptual models 'grounded in the data'.

Contents

Background

> 'In the country of the blind the one–eyed man is king.' — H.G. Wells

In his article in *Wired* Anderson (2008) proposed that in the age of the petabyte (10^{15} bytes) there was no longer any need for models or theories — in the sense that these were guesses, estimates, or hypotheses. Moreover, since statistical modelling or sampling of populations would be replaced by complete data sets, the scientific method itself was becoming obsolete. The data that could be assembled

ABOUT THE AUTHORS

Antony Bryant
Professor of Informatics, Leeds Metropolitan University

Uzma Raja
University of Alabama
Associate Professor, Management Information Systems

Program, University of Alabama

and made available for analysis would no longer be a sample, but would amount to a complete or near-complete population. Consequently the issues involved in moving from the findings from a sample to claims about the population in general would no longer be critical; and if data measured in petabytes proved to be insufficient then exabytes, zettabytes, and even yottabytes (10^{18}, 10^{21}, 10^{24} bytes respectively) would soon be upon us. (And if you are wondering what comes after these, then there are now newly coined terms including brontobytes and gegobytes, 10^{27} and 10^{30} bytes respectively.)

The influence of Anderson's article should not be under–estimated [1]. His earlier paper on 'The long tail' (Anderson, 2004) has proved highly influential, and the term itself has now entered the standard lexicon of many disciplines — particularly marketing, microfinance, business modelling, innovation, and social networking [2]. Consequently his later article was widely read, and evoked both critical responses and some measure of acceptance; continuing to be widely referred to in current papers on Big Data. Unfortunately it exemplifies a common misunderstanding of concepts such as 'model', 'theory' and 'scientific method'. The last term itself is a classic example of 'a contested concept' (Bryant, 2006; Gallie, 1956) very much up for grabs, since to be able to claim to be scientific is indisputably a good thing.

Anderson approvingly quotes George Box's dictum that 'all models are wrong, but some models are useful'. This is usually invoked to underline the idea that our knowledge of the world is based on approximations and interpretations, so it is more appropriate to judge these in terms of their usefulness, rather than their overall accuracy, however that might be assessed. Anderson, on the other hand, looks to the advent of Google — the specific company as well as the generic phenomenon — as exemplifying the process whereby models become obsolete: Once we have all the data there will be no more need for models, since they are at best incomplete. Anderson seems to assume that once all the data is collected it will be a fairly straightforward step to move seamlessly to 'correct', and presumably useful, conclusions via a range of computer–based and computational applications targeted on the data. This rather glosses over the manner and extent to which a complete set of data renders models or other forms of explanation obsolete: Anderson assumes this is the case, but there is the contrary argument — one with which the current authors would concur — that with a plethora of data there is often a greater demand for some form of model or abstraction. In part this is to avoid being overwhelmed by the detail of it all, but also because it needs to be understood that intrinsically explanation and understanding necessitate the use of forms of abstraction — *i.e.*, models which are developed in order to provide a focus on some aspects of the data, at the expense of others. Only in this way can any findings from the data actually be incorporated into our actions and strategies — *i.e.*, be of actual use: Something that Helles and Jansen (2013) stress in their editorial where they encourage responses to the 'invitation to scholarly and social dialogues about the data *that we all make as communicators, citizens, and consumers*' (stress added).

In taking issue with Anderson, therefore, we are implying that the terms 'models', 'theories', and 'scientific method' need to be better understood, both within specialized disciplines and in more general parlance. But before moving on to that it is important to recognize that Anderson's argument does encompass an important insight into the possibilities opened up with the dawning of what he termed 'The Petabyte Age', but which is now referred to as 'The Age of Big Data' — *i.e.*, the ability to derive patterns and correlations from huge data resources which can themselves prompt further investigation. In what follows we take Anderson's position as a starting point from which we seek to develop ideas about new opportunities and pitfalls for research and conceptualization. We offer a brief and highly specific overview of the history of scientific theory generation with a view to examining the ways in which the availability of massive data resources can be utilized to generate new theories: Affording an additional avenue for research and conceptualization, rather than a strategy that is seen as superseding the need for any models or theories at all.

The origins of the phrase 'The Age of Big Data' are unclear, but the term seems to have been used in the context of advances in computer technology in the 1990s. Diebold in his discussion of the term notes that '[T]he term Big Data, which spans computer science and statistics/econometrics, probably originated in lunch table conversations at Silicon Graphics in the mid–1990s, in which John Mashey figured prominently.' (quoted by Lohr, 2013)

N.B. The bulk of this paper was written in 2012, and draws on the example of Leetaru's use of Big Data which appeared in *First Monday* in 2011. But it also resonates with many of the themes and topics raised by the contributors to the Big Data issue of *First Monday* in October 2013; we have therefore incorporated some of the insights from these papers in what follows. In particular we see our paper as contributing to the opening up of 'the black boxes from which data can be seen to emerge' (Helles and Jensen, 2013).

Scientific method

Since at least the time of Plato and Aristotle there have been contending views about the sources of human knowledge, as well as about our relationship to the external world; respectively epistemology and ontology. In very simple and possibly simplistic terms, one of the products of the Enlightenment was the concept of the 'scientific method', as distinct from 'revelation'; with the latter's connotations of restricted access to knowledge and truth only being granted to specially chosen people or figures of authority. What actually constitutes the scientific method, however, is a highly fraught issue, so it is not too surprising that at certain times the Wikipedia entry for 'scientific method' contains the admonition 'editing of this article by new or unregistered users is currently disabled due to vandalism'. Foucault's term 'regimes of truth' is as applicable to post–Enlightenment times as it is to pre–Enlightenment ones, particularly in the Internet–cum–Wikipedia age.

On the one hand there have been those who take their lead from Aristotle and Francis Bacon who advocated the collection of large quantities of data, followed by exploration for patterns or regularities from which theories or hypotheses might then be derived, based on induction — *i.e.*, moving from a set of specific observations towards a more general conclusion. These hypotheses or theories could then be used as the basis for making predictions or deductions that themselves could be tested, thereby potentially corroborating or undermining the initial insights. But there have always been those who have questioned this seemingly non–controversial orientation. For instance Shapin and Schaffer (1985) offer an extended historical account in their analysis of the disagreement between Robert Boyle and Thomas Hobbes in the seventeenth century. Boyle's advocacy of public experiment was severely criticized by Hobbes who objected to the idea of arriving at truth by consensus — *i.e.*, those who attended the public experiments: Moreover a truth couched in probabilistic statements rather than definitive ones. In its place Hobbes argued in favour of certainty based on formally derived statements based on geometrical and logical foundations. Clearly in the long term Boyle's empiricism and experimental method won out, but the contentious nature of any attempts to generalize specific methods to a universal epistemology remains, particularly in the light of the onslaught that emerged in the later decades of the twentieth century from the writings of Karl Popper (1959), Thomas Kuhn (1970) and Richard Rorty (1979) amongst many others. Moreover there are many aspects of reality, both the social and natural, that are not readily accessible to 'experimentation'; *e.g.*, cosmology, climate science, and society in general.

Whatever one's position on the aforementioned issues, it is imperative to acknowledge these differences and dichotomies. Anderson fails to do this in his admittedly polemical and deliberately provocative article; particularly in his identification of scientific method with what, in another context, has been termed 'naive Baconian inductivism' — *i.e.*, the idea that stacking up vast amounts of data, observations or the like will lead to increasingly better (more complete, more certain, even definitive) knowledge. The weakness of this approach is something that should already be understood to some extent given the experience of the advent of such technologies as management information systems [MIS] in the 1960s and 1970s, which, it was claimed, would make management decision–making ever more effective, as more information was made available to the decision–makers. Already in the 1960s Russell Ackoff (1967) characterized and criticized this misconception in his classic paper 'Management misinformation systems', although a fairly mundane and uncritical view of 'rational behaviour' continues to be a common assumption in many areas.

The promise of Big Data

Anderson is, however, justified in directing attention to how best to appreciate and exploit the potential of these resources, and also to grasp that the developments encapsulated by the concept of Big Data may well have qualitative rather than simply quantitative ramifications. Consequently it is necessary to re–evaluate existing forms of analysis, and to encourage the development of novel and innovative tools and techniques for dealing with Big Data. But this is far from providing the basis for the jettisoning of tried–and–tested forms of reasoning and analysis. In concluding his article Anderson states that

> ... the opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?

Our views are broadly in line with those expressed by the respondents in the *Edge* discussion on Anderson's article (Dyson, *et al.*, 2008), seeing it as a mix of provocation, together with a mistaken view of the scientific method, but justifiably making the case for new forms of coping with and understanding the potentialities of massive collections of digitized data. Several of the respondents made the point that Anderson's arguments might have some application to marketing and advertising, but moving on from this to scientific theorizing was hardly warranted. Bruce Sterling summed it up as follows

> Yet I do have to wonder why — after Google promptly demolished advertising — Chris Anderson wants Google to pick on scientific theory. Advertising is nothing like scientific theory. Advertising has always been complete witch–doctor hokum. After blowing down the house of straw, Google might want to work its way up to the bricks (that's a metaphor). (Sterling, see Dyson, *et al.*, 2008)

Anderson himself does not seem to have engaged with his critics, but he seems to have largely reiterated his views in the keynote address he gave to the marketing organization conference, DMA 2012; albeit that the blogger reporting on the presentation notes that Anderson contrasted

> true data–driven, test obsessed companies with visionaries and innovators. He compared Amazon who, as king of the digital A/B test, drives its strategy and marketing almost wholly by data, to Apple at the other end of the spectrum. Anderson's view is that neither approach is right or wrong — without data, accurate decisions could never be made, whereas without innovation, new concepts would never be born. A dichotomy which is, or should be, close to every modern marketer's heart (Cummins, 2012).

So it would appear that Anderson is now offering a more nuanced view of Big Data, seeing innovations and insights coming from a variety of sources. Moreover in the interim the claims for and understanding of Big Data have developed, so that it is largely accepted that the skills needed to operate in this field include not just technical skills and expertise centred on analytic tools developed for specialist applications like astronomy (SKYCAT), fraud detection (HNC Falcon and Nestor PRISM), and financial transactions (various), but also the ability to present the outputs visually and also to understand the questions to pose in the first place. Consequently the argument that there is a growing and largely unresolved demand for 'data scientists' is continually countered by those who assert that '[E]nterprises won't need data scientists as their applications will process and analyse the data for them. Yes, someone will still need to know which questions to ask of the data, but the hard–core science of it should be rendered simpler by applications.' (Asay, 2013a)

We take up Anderson's argument and its aftermath, particularly developing the view that although Big Data offers significant opportunities, it does not preclude the necessity for insight and innovation; on the contrary, it makes new demands on people seeking to understand these opportunities — making good use of the potential benefits, and guarding against the pitfalls and hazards. For instance research indicates that someone's Facebook 'likes' can be used to predict a person's personality (Bachrach, *et al.*, 2012). Coincidentally, the day BBC reported on this research (12 March 2013) was also the date on which a report from Her Majesty's Inspectorate of Constabulary noted that a failure to share intelligence allowed a well–known celebrity to avoid arrest on a number of reports of sexual harassment — and worse — which were never linked (Casciani, 2013).

In the age of Google, an unprecedented amount of data is constantly amassed, and the figure grows on a daily basis — a phenomenon that is collectively of our own making, and that as Helles and Jensen stress is also constantly being processed and remade 'for different purposes'. Public and private sector organizations use intelligent agents to sift through data and make decisions with an aim to guiding policies or to maximize profits and sales. There is a constant move towards online migration of activities and resources, which threatens traditional institutions and media, such as education and newspapers respectively. Although initially this trend might appear to amount simply to a shift of content from one location to another — albeit with different formats for control and innovative protocols and processes — it quickly becomes evident that the ramifications are substantive and significant. Anderson takes this point to something of an extreme with his argument that we will no longer need theories, models, or hypotheses. In a related, but far more pessimistic view, Nicholas Carr (2008) argues that such developments lead to a dumbing–down, with people having shortened spans of attention, and generally becoming 're–programmed' into adjuncts to the Internet.

In the light of Anderson's and Carr's arguments it is worth considering Plato's *Phaedrus* in which Socrates discusses the new technology of writing. He explains that writing was discovered by the god Theuth, who presented it to King Thamus of Egypt saying that it would increase wisdom and memory. But Thamus demurred from this, explaining that, on the contrary, the contrivance of writing things down will result in people gaining a reliance on the written word at the expense of their memory. Moreover this will lead to a hierarchy, with the written word being superior to the spoken word; hence a diminished role for oratory and dialogue. Anderson's and Carr's contending positions can then be seen as twenty–first century equivalents of these respective arguments. Taken together, they correctly point to the potential and often unexpected or paradoxical ramifications of technological developments — echoing Socrates' discussion of the possible impact of technical innovation. In similar vein we argue that the era of Big Data offers new Opportunities for all of as we make our own contributions to Big Data, and as we also undertake research and analysis activities through the Internet. Such endeavours need to be understood and employed with some care: the example of the recent work on Facebook 'likes' exemplifying this double–edged aspect.

A brief note on theories

So far we have deliberately used the terms theory and model interchangeably. The terms themselves are ambiguous and the distinctions between them are vague, with differing arguments concerning the nature of their inter–relationship. Although it is often assumed that a model has a far more informal status than a theory. In general parlance the word theory can be understood as either something akin to a hunch — 'it's only a theory', 'it's just my theory' — or as referring to a well–structured and robust basis for explanation, prediction, and control — the theory of gravity, helio–centrism, planned behaviour, or plate tectonics would be candidates in this sense. The continuing efforts by those keen to argue that the universe is the product of some form of intelligent design are usually premised on the claim that the theory of evolution is only a theory; thereby implying the first of the two senses of the term, whereas opponents of this view would argue that it is very much a theory in the second sense.

The work of Mary Hesse (1963) has shed considerable light on the discussion of the terms theory and model, also the related concepts of metaphor and analogy. But for present purposes it is sufficient to note that our views are founded on John Dewey's (1930) pragmatism which views theories, models, concepts, or any other form of explanation in terms of usefulness rather than criteria pertaining to veracity or formality. In this sense theories and models are tools, designed to be used in specific contexts, rather than aiming to be over–arching explanations that apply across broad swathes of our existence. This resonates with Box's adage about models being judged in terms of use rather than completeness.

Critical issues for Big Data

As the claims for Big Data have been presented, they have been challenged by various critiques — for instance in the *Edge* discussion (Dyson, *et al.*, 2008). The accuracy and completeness of Big Data sets have been brought into question, with people pointing out that a significant proportion of the data is often missing or incomplete, and the data that is present can contain erroneous or ambiguous values [3]. In addition some of the stages of preparation and analysis are far from neutral and non–controversial. Despite this the insights and models that can be derived from Big Data have certainly been used effectively in the realms of advertising and marketing, justifying Anderson's point that 'Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day. And Google was right.' (Anderson, 2008) But to move from this to more grandiose claims about the end of science as we know it is somewhat far–fetched. In some respects these echo claims made about AI and expert systems since the 1970s, many of these based on the success of chess–playing programmes, with the assumption that all forms of knowledge and reasoning were similarly rule–based and accessible to computation.

It must also be acknowledged that Big Data has started to provide invaluable material for astronomers, epidemiologists, forensic scientists, and medical specialists particularly oncologists; but in all these cases it has only done so in conjunction with the specialist knowledge and expertise of the practitioners themselves — *i.e.*, the '–ists' not the '–isms'. In other words this conjunction of data, computer–based analytic tools, and the skills and insights of specially trained and experienced people has resulted, in some cases, in the development of new and improved theories and insights, enhanced levels of understanding, and more effective policies and interventions. In any massive set of data it will be possible to detect 'patterns', either through use of computational tools or from people

looking at the data in some way, but it is now understood that in some cases these patterns may be spurious. Indeed the term *apophenia*, originally coined with regard to various forms of neuroses, is now applied — in a non–clinical manner — to this phenomenon in the context of Big Data; referring to the detection of patterns where none exist [4].

Apophenia can be defined as the experience of seeing patterns or connections where there are none, and was originally seen as evidence of a mental disorder, particularly schizophrenia. John Nash, the central character in *A beautiful mind*, exhibits these symptoms. In less extreme cases it also applies to gambling, for instance where people study the sequence of numbers that come up in successive turns of a roulette wheel, believing that once they understand the pattern of results they can crack the system: in reality each spin of the wheel is entirely independent of preceding or succeeding ones. In the context of this discussion, the term represents a warning that in some cases the patterns that can be detected in Big Data may be illusory. On the other hand it may be possible to identify patterns from Big Data either through implementation of newly emerging techniques, or as a result of investigation by people with specific expertise or insight.

Big Data is then no basis for jettisoning expertise, nor is it justification for claiming that 'more is better'. Anderson may have achieved renown on the basis of his paper on 'The long tail', but the person often regarded as 'the father of the long tail' is Benoit Mandelbrot, and his work and career offer useful counterpoints to claims of the type that are made by Anderson and other proponents of Big Data. Mandelbrot is famous for his work on fractals and roughness, which provides the basis for many of the uses of massive data sets. Mandelbrot has recounted the episode in which, in 1961, he was on his way to give a seminar at Harvard

> I stepped into the office of my host, a Harvard economist. On his blackboard, I noticed a diagram nearly identical to one I was about to draw. His diagram referred to a topic of which I knew nothing: records of the price of cotton. My host had given up his attempt to model this phenomenon, and he challenged me to take over. (Mandelbrot, 1963)

This led him to write his now classic paper 'The variation of certain speculative prices' (Mandelbrot, 1963), which offered completely new ways of analyzing data — leading to the concepts of 'long tails', 'fat tails', fractals, and roughness. In an interview for *Edge* (Obrist, 2008), Mandelbrot wondered how things might have been different had someone cleaned the blackboard before he entered the room. He refers to Pasteur's apothegm to the effect that chance favours the prepared mind, but adds that 'I also think that my long string of lucky breaks can be credited to *my mode of paying attention*: I look at funny things and *never hesitate to ask questions*' (stress added). Thus we are not discouraging leaps that involve 'pattern recognition', but rather emphasizing that such insights need some basis in prior experience or expertise.

In drawing the contrast between Amazon and Apple, Anderson similarly seems to recognize the necessity not only for brute–force data analysis, but also for human insight (Cummins, 2012). Mandelbrot's cognitive leap exemplifies this, incorporating aspects of serendipity and abduction (see below). As such the age of Big Data continues to demand that all researchers and analysts engage in processes of modelling and theory generation, continually offering further bases for establishing rigor and relevance in their theoretical and conceptual development. In effect these data sets are another, and increasingly important, resource for developing our insights, rather than something that effectively displaces existing approaches: Indeed there is something of a renewed necessity in encouraging the skills and tools that might lead to these conceptual developments and outcomes.

In an age that is characterized not only by Big Data, but also by big communication and social networking, it is important to ensure that there is broad dissemination of the challenges and opportunities afforded by digitization. Hence we now turn to consideration of the concepts underlying the ways in which we might all make profound and critical contributions to and use of Big Data. In turning to the technical activities involved in Knowledge Discovery from Data (KDD) we wish to illustrate both the promise and the limitations of these sources of theoretical insights [5]. Moreover by drawing attention to the resonances and complementarities between KDD and the Grounded Theory Method (GTM) we can illustrate the ways in which 'data', 'theory', and 'knowledge' need to be understood; also 'induction', 'deduction', and 'abduction'. Ultimately we see this paper as a contribution to age–old questions around the issue of the source and nature of knowledge, and the status of knowledge claims: topics that have taken on a new resonance in the light of the growth of the internet and in particular in the age of Big Data.

The nature of data

Both KDD and GTM have close links with data [6]: For KDD data is regarded as something to be mined and explored — a search for associations. For GTM data is seen as the bedrock for developing theories, a process which must be 'grounded in the data'. The metaphorical imagery is similar, and, as will be explained below, somewhat troublesome and misleading. But the critical point is that both approaches adhere to the principle of close investigation and analysis of data as a core activity in theoretical development. Barney Glaser, one of the founding authors of GTM, contrasts this approach to what he has disparagingly termed 'immaculate conceptualization' — *i.e.*, plucking concepts out of thin air. This is something of an extreme position, and needs to be tempered with the realization that the history of scientific discoveries is littered with examples of what appear to be precisely this latter form of insight, as will be explained later.

At this stage the key point of note is that both approaches can claim compliance with the constraint that the validity of a theory is dependent on the quality of the data on which it is based; then subsequently by the ways in which any theoretical claims can garner strength and support based on testing and corroboration. This was one of the key points that Glaser and Anselm Strauss (1967) were at pains to make in their manifesto for GTM, *The discovery of grounded theory*.

The language in which many of the claims about KDD are expressed echoes those of GTM; often with the same ambiguous or misleading ramifications. The manifesto and founding text of GTM is Glaser and Strauss' book *The discovery of grounded theory*, but whether GTM is about discovery or interpretation has remained at the heart of GTM debate ever since; becoming particularly prevalent in the past decade with the articulation of a constructivist form of the method (Bryant and Charmaz, 2007). So too with the use of the term 'knowledge discovery through data mining', often described as the process of using data mining (DM) methods to 'extract knowledge' from massive data archives. In this regard DM is a component of the KDD process, providing the means and techniques 'to extract and enumerate patterns from the data according to the specifications of measures and thresholds, using databases together along with pre-processing, sub–sampling and transformations of the data' (Fayyad, *et al.*, 1996).

Although many standard texts portray the relationship between data and information as that between raw material and processed product, this 'chemical engineering metaphor' has long been subject to criticism emanating from a semiotic or semantic perspective that refuses to endorse the distinction between data and information expressed in many standard texts (see Bryant, 2006, for an extended critique). It is somewhat disappointing to note that despite the longstanding and serious nature of these forms of criticism, many scholarly papers and core texts remain stubbornly resistant even to acknowledging the contested nature of these claims. In her paper for the recent *First Monday* special issue on Big Data, Markham (2013) offers a refreshing critique of the term, invoking the work of Geertz who is quoted to the effect that 'what we call our data are really *our own constructions of other people's constructions* of what they and their compatriots are up to' (stress added), and Bowker who argued that the term 'raw data' is an oxymoron.

On the other hand the metaphorical resonances of the term 'data mining' may now in the digital age have some considerable 'grab' given the ways in which vast stores of data can be searched and analyzed using current technologies. Mining may well be a reasonable label for these processes, although it must be understood that only a limited range of such activities centres on data processing performed by silicon–based entities, and these must be initiated, guided, and supplemented by meaning–oriented actions of carbon–based entities. So it may be valid to argue that the existence of massive data sets affords additional enhancements and opportunities for analysis and investigation, and that to some extent standard statistical models that try to take account of sampling are not always relevant when one has the entire population at hand. But we should not lose sight of the enduring relevance of modelling, interpretation and investigative methods in general, all involving conscious activities and choices. As Jensen (2013) puts it: '[T]he question is who codifies what — and whom — with what consequences'.

Investigation and research have always employed a wide variety of resources including documents, observations and various forms of field studies and participative activities, now in addition we have access to the expanding panoply of digital resources, including e–mail, blogs, Facebook, and Twitter. The process of researching, however, cannot simply be a matter of amassing data until something almost magically emerges from it. Moreover the aim of investigation and research is not simply to report upon or transcribe reality, but to derive patterns and offer critical and innovative insights, including explanations aimed at responding to 'why?' and 'how?' type questions. This has not changed with the advent of the data age, but the glut of data certainly appears to offer new opportunities for research. Furthermore, a serious consideration of the nature of these new prospects can help to shed light on some of the general issues and features of the development of theoretical and conceptual insights.

An example of Big Data analytics, and its wider ramifications: *Culturomics 2.0*

In 2011 *First Monday* published a paper by Kalev Leetaru (2011) offering a contribution to what he terms the 'emerging field of Culturomics' [7], which 'seeks to explore broad cultural trends through the computerized analysis of vast digital book archives, offering novel insights into the functioning of human society'. Leetaru sees his work as heralding Culturomics 2.0, since he has taken this work a stage further, adding 'higher–level knowledge about each word, specifically focusing on news–tone and geographic location'. In September 2011, the BBC reported on his work using the highly misleading headline 'Supercomputer predicts revolution':

> A study, based on millions of articles, charted deteriorating national sentiment ahead of the recent revolutions in Libya and Egypt. While the analysis was carried out retrospectively, scientists say the same processes could be used to anticipate upcoming conflict. The system also picked up early clues about Osama Bin Laden's location. (BBC, 2011)

Leetaru cannot be held responsible for the headline, nor for the report itself, and the body of the news report does offer a far more nuanced account of the work, something that a close reading of Leetaru's paper confirms. On the other hand it is evident from Leetaru's own account that he clearly is making claims for the predictive capabilities of his approach, albeit that the analysis carried out was entirely retrospective. More importantly for our present purposes his overall strategy provides the basis for a discussion of the ways in which the analyses of and findings from Big Data need to be evaluated, critiqued, and challenged rather than received and accepted purely and simply as the result of applying non–controversial computational analysis to a massive set of data — in this case amounting to several million news articles, several billion words.

Leetaru's study builds on previous work in *Culturomics*, combining this with studies that investigated the 'tone' of news reports. His work is premised on the assertion that news reports 'contain far more than just factual details: an array of cultural and contextual influences strongly impact how events are framed for an outlet's audience, offering a window into national consciousness'. He characterizes his data domain in noting that 'accurately measuring the local press in nearly every country of the world requires a data source that continuously monitors domestic print, Internet, and broadcast media worldwide in their vernacular languages and delivers it as a uniform daily translated compilation'. Consequently the study uses the *Summary of World Broadcast* (*SWB*) collection, plus various other sources including the *New York Times* (*NYT*) digital archive. This data is then sampled by context (social media) and location (Egypt) — the latter being chosen as 'having the highest penetration of social media of any Middle Eastern or North African country'. Leetaru then goes on to offer an extended discussion of the advantages and weaknesses of the various data sources selected. The sources were then subjected to 'two key text mining techniques: sentiment mining ... and full–text geocoding'. These were used first on the *SWB* archive, then on the *NYT* source in order to provide two samples for comparison. The outcomes were then transformed into graphical figures, with supporting explanations in terms of changing level of tone with regard to specific terms.

Leetaru's method can be characterized in terms of KDD, a term which came into prominence in the 1990s, evolving from and encompassing the earlier process of data mining, and is now associated with terms including business intelligence and data analytics. As the issue of Big Data has developed, KDD has evolved in terms of associated tools and techniques, but essentially it encompasses the way in which massive data sources can be used as the basis from which to derive patterns and models, often with a commercial interest guiding the agenda — hence the link with business intelligence. The key features can be summarized as follows:

> Developing an understanding of the application domain
> Creating target data sets
> Data cleaning or pre–processing
> Data reduction and projection
> Data mining
> Interpretation of results

Developing an understanding of the application domain

Leetaru can be seen to have oriented his work around an interest in Culturomics wedded to a view that interrogation of large text archives provides a basis for 'insights to the functioning of society, including predicting future economic events'. [8] This is derived from his understanding of previous work — *e.g.*, his references to Bollen, *et al.* (2011), Gerbner and Marvanyi (1977), Michel, *et al.* (2011), and Mishne and Glance (2006). This then provides the basis for his use of similar data

sources and data mining techniques guided by a strategy focusing on tone and geographical analysis as possible constructs.

Creating target data sets

The target data sets are the *SWB* collection and then the *NYT* archive; Leetaru offers a fairly extended account of the rationale for choosing these, with analysis of the latter source being used as a check on the results emanating from the analysis of the former. He also used a 'Web crawl of English–language Web–based news sites' in order to ascertain the coverage of *SWB* with regard to non–Web sources.

Data cleaning or pre–processing

In KDD projects this step involves operations concerned with removal of outliers, deciding on strategies for missing values and identifying the sample population. Leetaru offers some account of this last aspect, and his Web crawl and later analyses using 'two well–known tonal dictionaries' can in part be seen as a strategy for coping with outliers and missing values, although there might be some doubts regarding his coverage of non–English and non–Web sources.

Data reduction and projection

This involves processes like data transformation and reduction of dimensionality, which may involve identification and creation of new variables using text mining. Leetaru's research side steps these aspects since it is driven by the initial strategy focused on tone and location. As reported in the paper, and elsewhere in his subsequent publications, the results seem to have dovetailed neatly with his initial suppositions around these two concepts; had they not then perhaps some forms of data reduction and projection might have been required.

Data mining

This step involves using various KDD techniques to search for patterns in the data and thus afford a basis for the creation of models. Such techniques include supervised methods (*e.g.*, regression, neural networks) or unsupervised methods (*e.g.*, clustering). The choice of technique depends upon the nature of the research question and the dataset. Leetaru used 'sentiment mining' and 'full–text geocoding'; in each case he offers an outline of the algorithmic basis of each technique.

Interpretation of results

Leetaru's paper is replete with many graphical representations of his findings, together with extensive discussion and explanations of the results. He also extends his work by taking his initial results and applying the model to other locations and time periods. Regardless of the actual techniques used in the previous steps, the process of interpretation is critical, both with regard to the ways in which Leetaru himself explains his findings, and the ways in which others then interpret and evaluate his work. In particular it has to be noted that he has continued to claim that his approach has predictive value, something that has been taken up and reported across the media and many other sources — something to which his Web site attests [9].

By briefly characterizing Leetaru's work against the steps of KDD, we can also raise the issue of how an alternative approach might have been used, in so doing exposing some of the hidden suppositions underlying many Big Data findings. In order to expand upon this further we will compare and contrast KDD with the Grounded Theory Method [GTM]. Both methods or approaches specifically centre around investigation of 'data', with the aim of producing useful conceptual outcomes: KDD is an approach designed for generating models from extensive digital data sources; GTM is a tried and tested research method that advocates an iterative process of data gathering and analysis in order to develop conceptual models or theories. What we intend to demonstrate is that lessons can be learned from both approaches with regard to analyses of Big Data — both in terms of undertaking the analysis itself and being able to take a critical view of other people's analyses.

In fact both KDD and GTM can be seen as instantiations of hermeneutics — the GTM approach of iteration between data gathering and analysis is akin to the hermeneutic circle in which our understanding of certain detailed aspects is dependent on our understanding of the whole, which is itself dependent on understanding the details. The aim of both GTM and KDD is best thought of not as a circle but rather as a spiral moving from an origin focused on the data towards ever higher levels of abstraction and conceptual reach. Unfortunately much of the hype and writing about Big Data fails to acknowledge this, instead usually presenting the findings as somehow definitive.

The grounded theory method — A very brief overview

The key precept of GTM is a constant engagement with the data. Bryant and Charmaz (2007) offer a succinct characterization of the method:

> Most fundamentally, grounded theory methods emphasize analyzing data and entail an iterative process of simultaneous

> data collection and analysis through which each informs and focuses the other. Thus, throughout the research process, grounded theorists engage in emergent analysis to direct their subsequent data collection, which, in turn, they use to develop and check their emerging ideas.

This should not be taken to mean that GTM is an 'inductive method' akin to the sort of endeavour evoked by Aristotle's idea of gathering a 'complete' set of data before starting to generate hypotheses. Certainly Glaser and Strauss appeared to propose it as 'inductive' and 'grounded in the data' in their early work, as they were determined to offer a distinct alternative to the form of hypothetico–deductive model which they saw as prevalent amongst social researchers in the U.S. in the 1960s. But the metaphor of 'discovery' is misleading, based on the same mistake that Anderson makes with comments along the lines of '[W]ith enough data, the numbers speak for themselves'. In both cases — GTM and analysis of Big Data: *It just ain't so.* Those involved ineluctably take on an active role as interpreters and analyzers, producing initial results that should then be used as the bases for further investigation and subsequent analyses. As will be explained below, a better term for GTM is *abductive.*

The key features of the method can be summarized under the following headings [10]:

> An initial interest in a problem domain or context
> An open but purposive sampling strategy in the earliest stages
> Simultaneous and iterative data collection and analysis
> Construction of various higher level abstractions — codes and categories in the parlance of GTM — derived from examination of the data, and not from previously derived theories of logical categories
> Repeated sampling and analysis in order to perform constant and repeated comparisons of the data in order to develop theoretical concepts and abstractions
> Selection of one or more specific concepts for further development
> Application of the selected concepts for use in a more deliberate manner against the context and appropriate data — theoretical sampling
> Articulation of theoretical statements and constructs offering a substantive account of some aspects of the initial context

We shall refer to all of these in what follows — offering an alternative approach for data–driven investigation and analysis, whether aimed at Big Data or more limited resources.

Getting started — Developing an understanding of the data domain

The various contributors to the October 2013 issue of *First Monday* made the point that although the term 'data' initially referred to what was given or present–at–hand, this is highly misleading as data is usually gathered or 'harvested' (Helles, 2013) with some specific aim in mind, and with regard to Big Data it is crucial to ensure that there is some level of clarity and comprehension of the agenda underlying the implementation of specific analytic tools and strategies to existing data sets that often appear simply to have *'grow'd like Topsy'*.

In the context of a formal research project one is expected to start with a hypothesis or detailed set of research questions; less formal investigative endeavours often start from an agenda or specific issue. But this is not the only possible stating point, since it implies that to some extent one already knows what the issues or problems actually are. As Albert Einstein put it 'If we knew what it was we were doing, it would not be called research, would it?' So in many instances investigation may well start from a position where one is unable to frame hypotheses or specific research questions or detailed agendas: *i.e.*, a hunch or the desire to follow up on some personal experience. Such was specifically the view of Glaser and Strauss in promulgating GTM. They sought to provide a justification for researchers keen to investigate a context or topic, however vague, without having to provide a detailed account of the nature of their investigation from the very beginning. Similarly KDD requires that in the early stages investigators seek an understanding of the problem domain, engaging in exploratory research aimed at articulating a research problem in such a way that useful variables can be identified from the dataset to formulate models. Although it can be contended that this is 'data fishing', it should also be understood that having too constrained an idea about hypotheses, research questions, or agendas early on in the knowledge discovery process may be inappropriate in some instances — a

stance that is very much the same as that proposed for GTM research. To extend the mining metaphor, it is perfectly possible to prospect a domain or data set with no clear idea of what might be mined, or at least with one's mind open to the possibility of finding unexpected results. Ultimately a clearer idea of what is involved in the data domain will be required, but this need not be obvious at the outset. Indeed it may be one of the early outcomes of the research exercise itself, just as it is in GTM when researchers move to theoretical sampling (see below).

Whatever the reason for starting the investigation, which may well be some personal motivation, the initial stage is to engage with the research domain and gather some data — possibly in the form of interviews, observations, documents, a large data set or a combination of all of these. Once this has been done to some extent, one can begin to start breaking the data up into 'incidents' or 'events' and comparing them with each other — Glaser and Strauss termed GTM the method of 'constant comparison'. This exploratory phase of 'open coding' leads to identification of issues in the sense that the researcher(s) begin to identify patterns across incidents, in much the same way as KDD practitioners look to identify useful variables from the dataset to formulate models. In both cases this is an activity to promote early forms of abstraction that can provide the basis for further investigation of the research domain in a more focused manner. In contrast to Leetaru's research — where he specifically interrogated his data with 'tone' and 'location' in mind — GTM advocates wide–ranging queries of the data in the first instance — *e.g.*, 'What is happening here?', 'What is this data a study of?', and so on. The researcher has to be prepared to be surprised by the results of such queries; keeping an open mind for possible outcomes.

Thus both KDD and GTM start from a grounded–ness in the data; albeit with the former usually oriented to a largely passive data set, and the latter usually associated with a context populated and constituted by social actors. But for each the basic issue is how can investigators make sense of large amounts of rich data without using preconceived strategies for classification and decomposition. This is not to say that such forms of preconception are misguided, but rather to stress that there are other useful and proven strategies worthy of consideration. The outcome of the earliest stages of GTM is a set of tentative categories or codes developed by the investigator(s) following analysis of an initial sample; the process is termed 'open coding' or 'initial coding'. The resultant codes, or sometimes an explicitly selected subset, are then used as the basis for further investigation, guided by 'theoretical sampling'; 'sampling for further theory construction to check and refine conceptual categories, not for representativeness of a given population' (Charmaz, 2007). This form of sampling can be contrasted with initial sampling, which is wide–ranging and contextually contingent. Theoretical sampling is performed with the aim of adding detail and coherence to the initial codes and theoretical categories under development.

On the basis of these insights from GTM and KDD, when confronted by Big Data — either as an investigator or as someone responding to someone else's findings — one should seek clarification regarding the nature of the data and the ways in which it has been used in the articulation of aspects of interest — *i.e.*, variables, patterns, codes or other abstractions. Although it may be countered that the GTM strategy may not be readily adapted to the massive data sets that now confront researchers, it is important to keep in mind that there are always other possible ways of interrogating a data set whether in the form of asking different questions or using a different algorithm.

Developing a focus

Once some initial level of abstraction has been accomplished, GTM researchers take a more targeted approach to data gathering. This is based on developing abstractions or concepts as a result of having accomplished the initial coding or categorization of the data — *i.e.*, the researcher(s) should now have some idea of what is going on in the context under investigation, and what the data can be understood to indicate. This more focused engagement with the data resonates with KDD where once the research problem has been identified or characterized a target dataset needs to be created or derived from an existing one. This will involve selecting variables of interest, usually in the form of keywords or phrases, that need to be identified from the data set, but this needs to be done in a controlled manner to avoid a proliferation of variables that may result in misleading or ambiguous results. Conversely too few variables can result in incomplete or incoherent models. The key point is that this stage necessitates conscious choices by the investigators to limit the scope and range of the investigation.

In both cases the approaches call for an informed and clearly stated set of selection criteria; with a cogent rationale for choosing to focus on one or more specific aspects within the target domain, and often explicitly excluding others. This newly derived basis then serves as a guide for further investigation in terms of identified variables, patterns or concepts. In this sense the subsequent stages of the analysis can be said in part to be 'driven by the data', but should not obscure the crucial role played by the active and hopefully insightful role of the investigators themselves. Indeed many of the key activities will be driven by the investigators as they move to re–evaluate the data at hand. This may include something along the lines of the KDD stage of data refinement, involving the

identification and bracketing or disregard of outliers, deciding on strategies for missing values and further characterization of the sample population: procedures designed to produce a more focused basis for analysis, based on detailed and iterative investigation of the data at hand.

There is some divergence here between GTM and KDD; for GTM the issue of missing values does not arise since the developing abstractions must be grounded in the data that has been used, not data that is 'missing'. The extent to which this is something of importance will depend on the context. Moreover in GTM there is a distinction between outliers and negative cases. Negative cases may provide a rationale for enhancing one's concepts to encompass varying aspects that can take account of the instances under consideration; hence a negative case may be taken account of by adding a new dimension to the analysis. On the other hand some investigators would argue that the robustness of one's findings depend in part on the extent to which negative cases or contrary findings were sought — along the lines of Popper's advocacy of a strategy of conjecture and refutation. Outliers on the other hand, or what Taleb (2010) has termed 'black swans', may require a major shift in focus that in essence leads to another investigatory project in itself.

In generic terms both KDD and GTM stress the necessity for moving on from the initial 'findings'; refining and narrowing the focus, on the basis of an iterative process of sampling and analysis. This may then necessitate reducing the dimensionality of the data, as well as creation of new variables: Text mining in KDD parlance, coding in various ways in GTM [11]. But the aim in all cases is to move from low level and manifold codes or variables towards more abstract ones with greater conceptual power and reach.

In effect these diverse strategies all centre on a search for patterns or other forms of regularity. KDD utilizes a range of methods and tools under the general heading of 'data mining'. These can be classed as supervised methods (*e.g.*, regression, neural networks) or unsupervised methods (*e.g.*, clustering). The choice of technique depends upon the nature of the research question and the data set. The activity of 'searching' for patterns, and the 'creation' of models both involve engagement with the data — resonating with the grounded–ness of GTM. What has to be understood is that these are all processes of active interpretation and interrogation of the data. Any suspicions of *apophenia* can be allayed by clear articulation of the bases for such findings, explanations and justifications that have to go well beyond simple–minded claims that the patterns are *really there in the data.*

Again we stress that the reason for dwelling on these aspects as they are incorporated into the two approaches is to provide a basis for the assessment and evaluation of the plethora of Big Data 'findings' that are already becoming a central feature of what can be termed 'familiar knowledge', emanating from Internet sources and searches, news reports, policy initiatives and other aspects of our time. It will be increasingly important that these findings and pronouncements are approached and evaluated with an understanding of their dependence on the skills, decisions, and choices of researchers and analysts in selecting appropriate application of techniques and methods. Moreover since findings of Big Data analyses are themselves forms of data, anyone seeking to understand or incorporate such accounts in any manner needs to be aware of the necessary skills involved in interrogating the data at hand, and the various ways in which we inevitably tend to categorize and stress particular features as we seek relevant and useful conceptual insights. In all such contexts it must be recognized that we are inevitably actively involved in an engagement with the data — a process of dialogue rather than one of discovery: Hence the importance of 'theoretical sensitivity' and abduction which we address below.

In the case of Leetaru's work, which we are using as a noteworthy exemplar of Big Data type research and not because we take issue with it in any way, we can see that from the outset he focused on the specific issues of tone and location. In so doing he was careful to evaluate possible data sources and explain the ways in which he carried out 'sentiment mining'. But this leaves open the possibility that he simply found what he was looking for. Perhaps a more open–ended search would have led in different directions or resulted in a more ambiguous set of findings. Thus instead of subjecting the entire data set from *SWB* to this form of analysis he might have taken a small sub–set and carried out a less constrained form of analysis, before proceeding to the full–blown data mining itself. Moreover it may be argued that although he is careful to describe the processes of sentiment mining and geocoding that he used, the actual algorithms incorporated in the tools themselves are somewhat opaque. Indeed the bulk of his paper is devoted to his own extended description and characterization of the findings themselves.

What his paper illustrates is that research using Big Data should be seen as affording the potential for developing new insights that requires a methodical approach consciously combining computational analyses of these resources with expertise of researchers and practitioners. We need to highlight the key issues and priorities in this new context of conceptual and theoretical development, recognizing that researchers may enter the process at different points and with different (tacit) perspectives. They may also enhance or develop the process by returning to more 'classic' research modes, or to other alternatives. In the case of Leetaru's work it

can be seen that what was been erroneously reported as a case of computer technology directly offering predictions is in fact a far more complex process derived initially from human ingenuity and suppositions, combined with the computational power and massive data sets now available to researchers. The outcomes are reliant on these new resources, in the sense that, if at all, it would only have been possible to achieve similar results to the same degree previously with vast amounts of time and effort, probably extending over several years. Moreover now that two years have passed since the initial paper was published it is not unfair to ask if the predictive nature of the work has actually been borne out — *e.g.*, did Leetaru himself or anyone using his model predict the events in Egypt in 2013 that led to the ouster of President Morsi? What predictions might emanate from the model with regard to the path of events in Egypt or Syria?

Theoretical sensitivity and abduction

Although there are differences in the approaches and justifications emanating from various sources on GTM — particularly Glaser (1992), Strauss and Corbin (1990), and Bryant and Charmaz (Bryant and Charmaz, 2007; Charmaz, 2007) — all are agreed that initial engagement with the data, followed by initial analysis and subsequent iteration between sampling and analysis are at the very heart of the method. One objective is to avoid what Glaser terms 'immaculate conceptualization'; *i.e.*, generation of concepts with no grounding in a research context. In the context of KDD researchers investigate large samples of data that can be examined with a view to identifying patterns and develop initial ideas about useful concepts. One of the skills needed in these sorts of investigation is the ability to discard some aspects of the data and to focus on others that might provide a basis for concept development and theoretical innovation. In GTM this skill is termed theoretical sensitivity. Glaser's (1978) characterization of the term, as being 'sensitive to theoretical issues while scrutinizing the data' hints at what is involved. But Reichertz (2007) makes the point even more forcefully in discussing the necessity for understanding theoretical sensitivity as a form of *abduction*, since the result is to bring together the logic of discovery with the logic of justification within the context of methodological considerations. By so doing it highlights a key issue that should be central for all researchers and investigators, but whose importance needs to be stressed particularly to those undertaking Big Data type research. Key aspects of such activities really do depend on the skills of the specific researcher, tools and methods alone are not sufficient, although they may be helpful or even necessary. Researching is not simply the case of collecting data or evidence, the researcher is a key factor in the research landscape, a link in the chain that reaches iteratively around data, codes, concepts, knowledge discovery, DM, and tentative theories. In fact Glaser's (1996) gerund perspective on GTM is important here with its stress on the activity of theoriz*ing* as opposed to theories *per se*; and it is also crucial who is doing the theorizing.

Ironically Glaser himself has continually distanced himself from the issue of who does the theorizing, as opposed to those who have articulated a constructivist approach to GTM — particularly Bryant and Charmaz in their separate and co–authored works (Bryant, 2002; Bryant and Charmaz, 2007; Charmaz, 2007). Glaser's view, with its stress on 'theories emerging from the data' resonates with Anderson and others who imply — perhaps unwittingly — that aggregating data will lead inexorably to new insights; and both perspectives efface the active role of researchers. If one uses the process–oriented terms — model*ing*, theoriz*ing*, research*ing* — rather than the simple noun forms, the issues come more readily in to focus: someone is engaging in these activities, and different people will come to the research domain or the realm of Big Data with different skills, expertise, experiences, and presuppositions. Again this resonates with Geertz: 'what we call our data are really *our own constructions of other people's constructions* of what they and their compatriots are up to' (Markham, 2013, stress added).

Grounding research concepts in the data lies at the heart of GTM, and Glaser reiterates that 'all is data' in many of his writings. To an extent this is valid, in the sense that researchers can treat anything as data — interviews, documents, observations, archives, publications, Web sites, and so on. But it is misleading if it is interpreted to mean 'data is all', a view that readily arises if the idea of theories emerging from the data is taken too literally — something that applies to Anderson, albeit in a different context. In fact the solution to the issue was encompassed in some aspects of GTM, particularly those that Strauss brought to the method — albeit tacitly (see Bryant, 2009) — drawing on his background in pragmatism; in particular the concept of abduction [12]. The term itself originates in the work of C.S. Peirce, who considered it to be the only way in which novel insights might arise. Charmaz offers a definition of abduction as follows

> 'a type of reasoning that begins by examining data and after scrutiny of these data, entertains all possible explanations for the observed data, and then forms

> hypotheses to confirm or disconfirm until the researcher arrives at the most plausible interpretation of the observed data.' [13]

Reichertz offers another:

> 'Something unintelligible is discovered in the data, and on the basis of the mental design of a new rule, the rule is discovered or invented and, at the same time, it also becomes clear what the case is. The logical form of this operation is that of abduction. Here one has decided (with whatever degree of awareness and for whatever reasons) no longer to adhere to the conventional view of things.' [14]

The idea of entertaining 'all possible explanations' is an intriguing one, and it takes on new resonances given the potential for computational analysis of massive data sets. Once we move away from the idea that Big Data will magically provide the one correct answer or model, we can surely entertain the possibility that it will allow a wide range of possible explanations and outcomes depending on the form of analytics and techniques that might be applied. Moreover the search for 'plausibility' is not something that can be done computationally, it relies on the theoretical sensitivity of researchers, although it also demands subsequent rigor in moving back to the data to see if the dramatic insight can be justified. Abduction thus incorporates what Glaser terms 'immaculate conceptualization', but in a more substantive and positive sense. Serendipity and the sort of leaps of insight referred to by Mandelbrot can now be understood as abductive, leading to the generation of further research to confirm or disconfirm the insights themselves.

A recent article about the situation in Syria can be used to illustrate our argument, in the light of Leetaru's work. In her recent article on Syria, Barbara Walter (2013) listed 'four things that President Obama should keep in mind as he considers the feasibility of pushing for a negotiated settlement in Syria':

1. Civil wars don't end quickly.
2. The greater the number of factions, the longer a civil war tends to last.
3. Most civil wars end in decisive military victories, not negotiated settlements.
4. Finally, the civil wars that end in successfully negotiated settlements tend to have two things in common. First, they tend to divide political power amongst the combatants based on their position on the battlefield. ... Second, successful settlements all enjoy the help of a third party willing to ensure the safety of combatants as they demobilize.

Offering a cogent argument for this summation, Walter draws not on Big Data but on her significant expertise, and that of her colleagues. Certainly this is based on 'data' in the sense that it is probably derived from extensive studies of previous, similar circumstances, but it is hardly likely to have been arrived at simply by application of Big Data analytics. Indeed the contrast between Leetaru's approach and that of domain experts such as Walter echoes the debates that took place in the 1990s around the nature of artificial intelligence [AI]. The proponents of AI saw intelligence as essentially rule–based, so the technology required for a machine to exhibit intelligence centred on a vast and ever–growing rule base amenable to rapid processing whenever responses to specific questions and interrogations were demanded. This 'strong' programme for AI drew on the work of Herbert Simon and was the target of several trenchant critiques, including that of Dreyfus and Dreyfus (1986). In contrast to the rule–based idea of intelligence they argued that 'intelligence' was not a unitary phenomenon that simply developed in linear fashion, with 'more' inevitably being 'better'. Instead they saw a path from 'novice' to 'expert' as one of moving from a somewhat mechanistic rule–based practice to a far more insightful and seemingly intuitive basis, albeit one that still drew on experience and clear reasoning.

Thus a novice nurse is taught 'how to read blood pressure, measure bodily outputs, and compute fluid retention, and is given rules for determining what to do when those measurements reach certain values' [15]. A basic syntax is imparted to the learner, but it is seen largely as an admixture of independent rules. Later stages — advanced beginner, competence, and then proficiency — move on from this strict and narrow adherence to specific rules, gradually encompassing a wider and more diverse attempt to understand and respond appropriately to specific contexts, each with its unique characteristics. The highest level, expert, is reached when normal practice can largely be carried out with little or no recourse to specific rules; something that is only affected when novel or complex situations are encountered. The expert can function without recourse to conscious, analytic reasoning; the skills are intuitive, instinctive, automatic. Dreyfus and Dreyfus offered a dramatic illustration of this with regard to an expert chess player being given the task of adding numbers spoken to him at a rate of one number per second, while playing five–seconds–a–move chess against a player of only slightly lesser ability, and convincingly beating his opponent.

In essence this concept of 'expertise' is the embodiment of abduction; the ability to conceptualize and act upon insights which are derived neither from induction nor deduction, although if justification is required the expert should be able to explain their reasoning on the basis of one or other, or both, induction or deduction. Thus Walter offers links to various scholarly papers for each of the four aspects she raises; resources that refer to previous examples (a basis for induction) as well as to existing and tested models or theories (a basis for deduction and prediction).

Abduction then refers to a process of reasoning that relies on the intuition of someone with a claim to expertise — including detectives, real or fictional, with Sherlock Holmes providing the earliest and most notable example. Other examples may be drawn from medical practice, particularly in the realm of differential diagnosis; where patients can exhibit similar symptoms but in fact are suffering from very different ailments, and ascertaining the correct intervention is dependent on the skill of the practitioner. Again a lesson can be drawn from the world of AI in the 1980s and 1990s where specific implementations of AI techniques, referred to as 'expert systems', were used in various fields, including diagnoses of patients presenting with stomach pains. At the time it was reported that these expert systems were better at diagnosis than doctors, but on a closer reading it transpired that the main reason for this was that 'better' was simply seen in terms of the percentage probability of providing the correct diagnosis. In fact a significant proportion of patients presenting with stomach pains are likely to be suffering from appendicitis, so if this is offered as the diagnosis, then it is likely to produce a high level of correct hits. In fact most doctors understand this, but are often keen to rule out other, perhaps more acute or potentially fatal possibilities, hence producing a lower rate of correct diagnoses the first time round. A recent paper makes a similar claim, on a similar basis (Dvorsky, 2013), and some of the comments offer similar criticisms of the claims made. It would surely be preferable to receive diagnosis from an experienced doctor, possibly aided by technology in some form, than simply relying on the output from the computer. Similarly in terms of deliberating upon possible strategies and interventions in contexts such as Syria in 2014, consideration of the views of Walter and her experienced colleagues affords at least as feasible and fruitful a basis as does the work of Leetaru. Perhaps the work of both could be used in concert to provide further guidance and insight into a highly fraught and seemingly paradoxical context.

In the context of GTM abduction has taken on a renewed resonance particularly with the development of the constructivist account of the method, since it specifically moves away from the idea that concepts or theories 'emerge' in some fashion from the data, instead putting the onus fairly and squarely on the shoulders of the researcher(s). This then allows for the fact that in many instances researchers will carefully gather data, but may well then make significant conceptual leaps, which although evoked by the data cannot simply be explained from the data (see Blaikie, 2004, 2007); further analysis and data gathering will be required to substantiate these insights. This is what is partially encapsulated by the term 'theoretical sensitivity'. As such this conjunction of abduction with theoretical sensitivity is in direct contrast to the concept of apophenia referred to earlier: The former affirms the importance of researchers as active and insightful agents, investigating contexts and data; the latter indicates the possibility that the patterns or regularities that have been detected cannot be substantiated in any clear and reasonable manner. On the other hand, it must be understood that it may well be worth taking the risk of being accused of apophenia in order to offer innovative insights into the data at hand.

Although the term abduction is somewhat problematic (Plutynski, 2011; Minnameier, 2010), given that Peirce himself offered seemingly contradictory or divergent accounts of the term at different phases of his writing, it is important to recognize that, however diverse, it offers an alternative to the usual recourse to reasoning by induction or deduction. As Jensen (2008) points out, induction and deduction themselves are not immune to significant criticism, and it is only with abduction that the 'interchange between researcher and informants [can serve to] establish — infer — relevant categories and concepts'. Thus in the context of understanding and evaluating the findings of analyses of Big Data, abduction is required at least as much as any other forms of reasoning and justification.

Conclusions

What all of this demonstrates is that extending data analytics into more and more realms of our daily existence is a highly complex, social phenomenon. It does not reduce the need for insight and careful, robust research; on the contrary it adds to the issues that need addressing, and some of these are encompassed by comparing and contrasting approaches such as KDD, considered in general terms, with some of the key aspects of GTM — including theoretical sensitivity and abduction.

Although the term 'the age of Big Data' has been much hyped, this is not to deny that there have been numerous accounts of the effective use of

the outcomes of these new forms of analysis. The existence of extensive data archives, and the development and refinement of KDD tools and techniques, affords researchers new opportunities for analyses and theoretical insights derived from massive data sets. Some of these approaches have already yielded significant outcomes, particularly with regard to practices of marketing and advertising. In such contexts, however, the competitive edge that might be obtained may be only temporary. For instance the use of data analysis in the recruitment of baseball players, as recorded in *Moneyball* — the book (Lewis, 2003) and the film — was of maximum effectiveness while only one team knew of its capabilities. Another notable example was in the context of President Obama's re–election campaign (Issenberg, 2012), something that will surely be mimicked by all candidates with the necessary resources in the future.

In fact the Obama campaign did not simply crunch the data, but relied on people engaging directly with the data, and subsequently developing some initial insights and intuitive leaps to get this going.

> In late spring, the backroom number crunchers who powered Barack Obama's campaign to victory noticed that George Clooney had an almost gravitational tug on West Coast females ages 40 to 49. The women were far and away the single demographic group most likely to hand over cash, for a chance to dine in Hollywood with Clooney — and Obama.
>
> So as they did with all the other data collected, stored and analysed in the two–year drive for re–election, Obama's top campaign aides decided to put this insight to use. (Scherer, 2012)

More recently a report from the Executive Office of the President dating from March 2012 lists more than 50 Federal programs 'that address the challenges of, and tap the opportunities afforded by, the Big Data revolution'. The document itself offers brief accounts of these projects, mostly couched in terms of how these forms of analysis 'could enable' particular activities in the future. But it must also be understood that such findings can also act as self–fulfilling or self–falsifying prophecies. For instance Leetaru's findings regarding the tone of news reports may lead to some reporters 'toning up' or 'toning down' their reports in the light of the analysis to which they think their stories will be subjected. Similarly users of social media may purposely make liberal use of the terms in Leetaru's model in order to gain more coverage for their particular interests.

The issues of data ownership, control, and access are important and often evaded or obscured in reports on the uses of Big Data. Many of the commercial uses, particularly those by Google, Amazon, Facebook and the like, rely on data that is commercially sensitive and valued by the host organization. Moreover even those data sets that are more widely available are often only open to those with access to specific institutional facilities — *e.g.*, universities or research centres. Thus boyd and Crawford argue that there now exists a 'class of the Big Data rich ... reinforced through the university system: top–tier, well–resourced universities ... able to buy access to data, and students from the top universities are the ones most likely to be invited to work within large social media companies.' [16]

The algorithms underlying much of the power of Big Data have also come under increasing scrutiny. The recent furore regarding the marketing of 'Keep Calm ...' t–shirts brought the concept of algorithms into the wider realm, with the supposition that many aspects of Internet marketing were no longer within the remit of human responsibility (McVeigh, 2013):

> Amazon was forced to take action on Saturday after it was found to be selling T–shirts with slogans promoting rape and violence on its Web site.
>
> The American clothing company Solid Gold Bomb blamed an automated computer dictionary for its series of the items emblazoned with offensive phrases such as "Keep Calm and Rape a Lot" and "Keep Calm and Hit Her", based on the much reproduced "Keep Calm and Carry On" Second World War poster.
>
> Both companies were bombarded with complaints and Solid Gold Bomb later closed its Twitter account. The T–shirts were still on sale in Germany on Saturday.
>
> One of the slogans was taken down but others, including "Keep Calm and Knife Her"

> and "Keep Calm and Punch Her" remained on Solid Gold Bomb's Web site on Saturday afternoon; the company said they were all in a "deletion queue".

In fact in the case of the offensive t–shirts this was clearly not the case, since it was not possible to order a t–shirt bearing the slogan 'Keep Calm and Hit *Him*!' But this is not to undermine the importance of having some scrutiny of the algorithms underlying analysis of massive data sets. In his column for the *Observer* in the U.K., John Naughton (2012) has written about the secret power of algorithms, ranging from the relatively innocuous Amazon suggestions for what you may also like to buy, to the more potent ones that govern investment decisions and policy initiatives. He mentions the work of Nick Diakopoulos, who has analyzed the criteria underlying the Google page–rank algorithm — something that research such as that by Leetaru needs to take into account, with regard to the ways in which news is reported, categorized, and prioritized. Astrid Mager (2012) takes her analysis further, claiming that there is an 'algorithmic ideology' in effect, influencing the outcomes of online searching and modeling.

The power of many of these algorithms relies on them being unknown — Google's algorithm is one key example. If Web developers can get to know the underlying algorithm that Google uses, then they can tune their Web pages to ensure that they come at the top of the most relevant — usually commercial important — search queries. But if Google guards the algorithm from scrutiny, then the suspicion arises that they may be skewing the results in some particular manner to favour themselves or some other enterprises. The problem is that once the results are available to others, then people will have the opportunity to alter their behaviour accordingly.

Big Data thus suffers from a paradox of dissemination; not only in terms of the findings becoming self–fulfilling or self–falsifying prophecies, but also with regard to the algorithmic bases of the analytic tools upon which KDD is built. This is in addition to the growing concern with the ways in which the data is actually 'harvested' and then used. Issues around uses of people's Facebook 'likes' have now been dwarfed by the continuing revelations around PRISM (Wikipedia, 2013).

A recent paper by Asay (2013b) offers a well–balanced view on Big Data, offering an argument that outlines the way in which the 'phenomenon' or 'fad' can perhaps mature into a discipline, with less hype and a greater understanding of its potential uses, weaknesses, and complexities. What we have sought to do is provide a mapping of the theory and model building process in GTM with KDD to argue that the latter can be understood as a theory generation tool in the age of Big Data if researchers take account of a variety of methods that can foster conceptual innovation, specifically those encompassed by GTM. It is important to understand that we are not in any way suggesting that use of KDD, informed and articulated by GTM, is the only way in which theorizing can be accomplished; But it is a route worthy of consideration, and one that could contribute to the development from fad to discipline.

We must also point out that there is a major paradox at the heart of discussions about Big Data. In 2008 Anderson talked of 'the age of the petabyte', and we have now moved, in just a few years, to several orders of magnitude beyond this. All of which implies that for the foreseeable future the volume of data being produced will continue to grow in a similar fashion. If this is the case, then by definition we can never have a complete set of data; on the contrary the search for complete data is something akin to a digital task of a Sisyphus — we are forever doomed to gather complete data sets, only to see yet more data arriving by the millisecond. So the necessity for insight, modelling, and theorizing is ever upon us.

The quote at the start of this paper is taken from a short story by H.G. Wells in which a mountaineer has an accident that leads him to discover The Valley of the Blind. Since he has sight he assumes he will be able to rule over them, but instead discovers that the inhabitants of the valley cannot understand what sight involves and why it is necessary, since they are perfectly able to operate with their other senses. In similar fashion those extolling the unalloyed virtues of The Realm of Big Data need to consider the potential benefits of a similar learning exercise.

About the authors

Antony Bryant is Professor of Informatics at Leeds Metropolitan University.
E–mail: a [dot] bryant [at] leedsmet [dot] ac [dot] uk

Uzma Raja is Associate Professor in the Management Information Systems Program at the University of Alabama.
E–mail: uraja [at] cba [dot] ua [dot] edu

Notes

1. Earlier drafts of this paper elicited several responses that pointed out that Anderson's arguments were no more than a 'straw-man' — hardly worth bothering with; yet he has clearly set the tone for a core theme in the claims and hype around Big Data. As such we feel it important to engage with the ramifications and resonances of what may well have been a deliberate provocation — the straw–man has proved to to be something of substance and durability. Indeed a recent article (Steadman, 2013) focusing on Leetaru's work 'Big Data and the death of the theorist' demonstrates that the term itself continues to have some currency.

2. Helles refers to Anderson's paper, although making a different point with regard to the concept of The Big Tail.

3. Papers comprising *First Monday's* October 2013 special issue (http://firstmonday.org/issue/view/404) take up many of these issues.

4. Various authors, such as boyd and Crawford, have warned of this possibility in the context of Big Data, and it is interesting to note that many of those heralding the age of big data tend to offer analyses that indicate how patterns have been detected that relate to events in the past — *e.g.*, Leetaru's work. As more predictions and strategies are made based on big data analytics, examples of apophenia and false correlation will no doubt abound, but in the meantime Silverman's (2014) perceptive review of a recent encomium to big data is worth noting.

5. We use KDD to include data mining, machine learning, decision support systems; all of which can be used to derive knowledge from data sets.

6. The term data is treated differently on different sides of the Atlantic — in the U.S. the term is treated as a plural (hence 'data are ...'), in the U.K. it is used as a singular ('data is ...'). In this paper we have followed the U.K. usage.

7. There seems to some confusion regarding this term which is sometimes presented as *Culturnomics* — see http://nationalsecurityzone.org/war2-0/kalev-leetaru-on-culturnomics/ for a recent interview with Leetaru. The authors and readers might wonder if both terms would be picked up by a text–mining algorithm!

8. Leetaru seems to accept that predicting future economic events has been a successful venture — many would disagree with such a claim.

9. His Web site reports on many of these aspects; see http://www.kalevleetaru.com.

10. There is an extensive and growing literature on GTM — a good starting point for those unfamiliar with the method can be found in Charmaz (2006).

11. It should be noted that there is some variance in coding strategies amongst different GTM approaches.

12. In what follows only a very limited and specific account of abduction is offered. Readers should refer to the entry in the *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/entries/abduction/), also the references to Blaikie, Jensen, and Reichertz.

13. Charmaz, 2007, p. 188.

14. Reichertz, 2007, p. 219.

15. Dreyfus and Dreyfus, 1986, p. 22.

16. boyd and Crawford, 2012, p. 674.

References

Russell L. Ackoff, 1967. "Management misinformation systems," *Management Science*, volume 14, number 4, pp. B147–B156.

Chris Anderson, 2008. "The end of theory: The data deluge makes the scientific method obsolete," *Wired*, volume 16, number 7, pp. 106–129, and at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory, accessed 22 January 2014.

Chris Anderson, 2004. "The long tail," *Wired*, volume 12, number 10, and at http://www.wired.com/wired/archive/12.10/tail.html, accessed 22 January 2014.

Matt Asay, 2013a. "Data scientists: Do they even exist? Data data everywhere, but not a drop to shrink," *The Register* (13 February), at http://www.theregister.co.uk/2013/02/13/open_and_shut/, accessed 22 January 2014.

Matt Asay, 2013b. "Big data myths give way to reality," *readwrite* (26 December), at http://readwrite.com/2013/12/26/big-data-myths-reality, accessed 22 January 2014.

Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli and David Stillwell, 2012. "Personality and patterns of Facebook usage,"

WebSci '12: Proceedings of the 3rd Annual ACM Web Science Conference, pp. 24–32.
doi: http://dx.doi.org/10.1145/2380718.2380722, accessed 21 January 2014.

BBC, 2011. "Supercomputer predicts revolution" (9 September), at http://www.bbc.co.uk/news/technology-14841018, accessed 22 January 2014.

Norman Blaikie, 2007. *Approaches to social enquiry*. Second edition. Cambridge: Polity.

Norman Blaikie, 2004. "Abduction," In: Michael S. Lewis–Beck, Alan Bryman and Tim Futing Liao (editors). *Sage encyclopedia of social science research methods*. volume 1. Thousand Oaks, Calif.: Sage.
doi: http://dx.doi.org/10.4135/9781412950589, accessed 21 January 2014.

Johan Bollen, Huina Mao and Xiao–Jun Zeng, 2011. "Twitter mood predicts the stock market," Journal of Computational Science, volume 2, number 1, pp. 1–8.
doi: http://dx.doi.org.proxy.cc.uic.edu/10.1016/j.jocs.2010.12.007, accessed 21 January 2014.

danah boyd and Kate Crawford, 2012. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society*, volume 15, number 5, pp. 662–679.
doi: http://dx.doi.org.proxy.cc.uic.edu/10.1080/1369118X.2012.678878, accessed 21 January 2014.

Antony Bryant, 2009. "Grounded theory and pragmatism: The curious case of Anselm Strauss," *FQA*, volume 10, number 3, at http://www.qualitative-research.net/index.php/fqs/article/view/1358/2850, accessed 22 January 2014.

Antony Bryant, 2006. *Thinking "informatically": A new understanding of information, communication, and technology*. Lewiston, N.Y.: Edwin Mellen Press.

Antony Bryant, 2002. "Re–grounding grounded theory," *Journal of Information Technology Theory and Application*, volume 4, number 1, pp. 25–42, and at http://aisel.aisnet.org/jitta/vol4/iss1/7/, accessed 22 January 2014.

Antony Bryant and Kathy Charmaz, 2007. "Introduction: Grounded theory research: Methods and practices," In: Antony Bryant and Kathy Charmaz (editors). *SAGE Handbook of Grounded Theory*. London: Sage.
doi: http://dx.doi.org/10.4135/9781848607941, accessed 21 January 2014.

Nicholas Carr, 2008. "Is Google making us stupid?" *Yearbook of the National Society for the Study of Education*, volume 107, number 2, pp. 89–94.
doi: http://dx.doi.org/10.1111/j.1744-7984.2008.00172.x, accessed 21 January 2014.

Dominic Casciani, 2013. "The missed chances to get Jimmy Savile," *BBC News* (12 March), at http://www.bbc.co.uk/news/uk-21756150, accessed 22 January 2014.

Kathy Charmaz, 2007. "Constructionism and grounded theory," In: James A. Holstein and Jaber F. Gubrium (editor). *Handbook of constructionist research*. New York: Guilford Press, pp. 397–412.

Kathy Charmaz, 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. London: Sage.

D. Cummins, 2012. "DMA 2012: Big Data and the emergence of the long tail," *Marketsoft* (16 October), at http://www.marketsoft.com.au/2012/10/16/dma-2012-big-data-and-the-emergence-of-the-long-tail/, accessed 22 January 2014.

John Dewey, 1930. *The quest for certainty: A study of the relation of knowledge and action*. London: G. Allen & Unwin Ltd.

Francis X. Diebold, 2012. "A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline," at http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf, accessed 22 January 2014.

George Dvorsky, 2013. "Computers are better at diagnosing and treating patients than doctors," *io9* (13 February), at http://io9.com/5983991/computers-are-better-at-diagnosing-and-treating-patients-than-doctors, accessed 22 January 2014.

George Dyson, Kevin Kelly, Stewart Brand, W. Daniel Hillis, Sean Carroll, Jaron Lanier, Joseph Traub, John Horgan, Bruce Sterling, Douglas Rushkoff, Oliver Morton, Daniel Everett, Gloria Origgi, Lee Smolin and Joel Garreau, 2008. "On Chris Anderson's The End of Theory," *Edge* (30 June),

at http://www.edge.org/documents/archive/edge249.html, accessed 22 January 2014.

Usama M. Fayyad, Gregory Piatetsky–Shapiro and Padhraic Smyth, 1996. "From data mining to knowledge discovery: An overview," In: Usama M. Fayyad, Gregory Piatetsky–Shapiro, Padhraic Smyth and Ramasamy Uthurusamy (editors) *Advances in knowledge discovery and data mining.*. Menlo Park, Calif.: AAAI Press, pp. 1–36.

W.B. Gallie, 1956. "Art as an essentially contested concept," *Philosophical Quarterly*, volume 6, number 23, pp. 97–114.

George Gerbner and George Marvanyi, 1977. "The many worlds of the world's press," *Journal of Communication*, volume 27, number 1, pp. 52–66.
doi: http://dx.doi.org/10.1111/j.1460-2466.1977.tb01797.x, accessed 21 January 2014.

Barney G. Glaser (editor), 1996. *Gerund grounded theory: The basic social process dissertation*. Mill Valley, Calif.: Sociology Press.

Barney G. Glaser, 1992. *Basics of grounded theory analysis: Emergence vs. forcing*. Mill Valley, Calif.: Sociology Press.

Barney G. Glaser, 1978. *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, Calif.: Sociology Press.

Barney G. Glaser and Anselm L. Strauss, 1967. *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Rasmus Helles, 2013. "The big head and the long tail: An illustration of explanatory strategies for big data Internet studies," *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4874/3753, accessed 22 January 2014.
doi: http://dx.doi.org/10.5210/fm.v18i10.4874, accessed 22 January 2014.

Rasmus Helles and Klaus Bruhn Jensen, 2013. "Introduction to the special issue: 'Making data — Big data and beyond'," *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4860/3748, accessed 22 January 2014.
doi: http://dx.doi.org/10.5210/fm.v18i10.4860, accessed 22 January 2014.

Mary B. Hesse, 1963. *Models and analogies in science*. *Newman history and philosophy of science*, volume 14. London: Sheed and Ward.

Sasha Issenberg, 2012. "How President Obama's campaign used big data to rally individual voters," *Technology Review* (19 December), at http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/, accessed 22 January 2014.

Klaus Bruhn Jensen, 2013. "How to do things with data: Meta–data, meta–media, and meta–communication," *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4870/3751, accessed 22 January 2014.
doi: http://dx.doi.org/10.5210/fm.v18i10.4870, accessed 22 January 2014.

Klaus Bruhn Jensen, 2008. "Deduction vs. induction vs. abduction," In: Wolfgang Donsbach (editor). *International Encyclopedia of Communication*. Volume 3. Malden, Mass.: Blackwell Science, pp. 1,188–1,192.

Thomas S. Kuhn, 1970. *The structure of scientific revolutions*. Second edition, enlarged. Chicago: University of Chicago Press.

Kalev H. Leetaru, 2011. "Culturomics 2.0: Forecasting large–scale human behavior using global news media tone in time and space," *First Monday*, volume 16, number 9, at http://firstmonday.org/article/view/3663/3040, accessed 22 January 2014.

Michael Lewis, 2003. *Moneyball: The art of winning an unfair game*. New York: W.W. Norton.

Steve Lohr, 2013. "The origins of 'big data': An etymological detective story," *New York Times* (1 February), at http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/, accessed 22 January 2014.

Tracy McVeigh, 2013. "Amazon acts to halt sales of 'Keep Calm and Rape' t–shirts," *The Observer* (2 March), at http://www.theguardian.com/technology/2013/mar/02/amazon-withdraws-rape-slogan-shirt, accessed 22 January 2014.

Astrid Mager, 2012. "Algorithmic ideology: How capitalist society shapes search engines," *Information, Communication & Society*, volume 15, number 5, pp. 769–787.
doi: http://dx.doi.org/10.1080/1369118X.2012.676056, accessed 22 January 2014.

Benoit Mandelbrot, 1963. "The variation of some other speculative prices," *Journal of Business*, volume 36, number 4, pp. 394–419.

Annette N. Markham, 2013. "Undermining 'data': A critical examination of a core tem of scientific inquiry," *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4868/3749, accessed 22 January 2014.
doi: http://dx.doi.org/10.5210/fm.v18i10.4868, accessed 22 January 2014.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden, 2011. "Quantitative analysis of culture using millions of digitized books," *Science*, volume 331, number 6014 (14 January), pp. 176–182.
doi: http://dx.doi.org/10.1126/science.1199644, accessed 22 January 2014.

Gerhard Minnameier, 2010. "The logicality of abduction, deduction and induction," In: Mats Bergman, Sami Paavola, Ahti–Veikko Pietarinen and Henrik Rydenfelt (editors). *Ideas in action: Proceedings of the Applying Peirce Conference*, pp. 239–251, and at http://www.nordprag.org/nsp/1/Minnameier.pdf, accessed 22 January 2014.

Gilad Mishne and Natalie S. Glance, 2006. "Predicting movie sales from blogger sentiment," *AAAI Symposium on Computational Approaches to Analysing Weblogs AAAI–CAAW*, page 155–158, and at http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-030.pdf, accessed 22 January 2014.

John Naughton, 2012. "How algorithms secretly shape the way we behave," *The Observer* (15 December), at http://www.theguardian.com/technology/2012/dec/16/networker-algorithms-john-naughton, accessed 22 January 2014.

Hans Ulrich Obrist, 2008. "The father of long tails: Interview with Benoit Mandelbrot," *Edge*, at http://www.edge.org/3rd_culture/obrist10/obrist10_index.html, accessed 22 January 2014.

Anya Plutynski, 2011. "Four problems of adbuction: A brief history," *HOPOS: The Journal of the International Society for the History and Philosophy of Science*, volume 1, at http://155.97.32.9/~plutynsk/HOPOSproofs.pdf, accessed 22 January 2014.

Karl Popper, 1959. *The logic of scientific discovery*. London: Hutchinson.

Jo Reichertz, 2007. "Abduction: The logic of discovery in grounded theory," In: Antony Bryant and Kathy Charmaz (editors). *SAGE Handbook of Grounded Theory*. London: Sage, pp. 214–228.
doi: http://dx.doi.org/10.4135/9781848607941, accessed 21 January 2014.

Richard Rorty, 1979. *Philosophy and the mirror of nature*. Princeton, N.J.: Princeton University Press.

Michael Scherer, 2012. "Inside the secret world of the data crunchers who helped Obama win," *Time.com* (7 November), at http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/, accessed 22 January 2014.

Steven Shapin and Simon Schaffer, 1985. *Leviathan and the air–pump: Hobbes, Boyle, and the experimental life*. Princeton, N.J.: Princeton University Press.

Jacob Silverman, 2014. "Through a data set, darkly," *Pacific Standard* (8 January), at http://www.psmag.com/navigation/books-and-culture/data-set-darkly-quantitative-analysis-secret-understanding-culture-72410/, accessed 22 January 2014.

Stanford Encycopedia of Philosophy, 2011. "Abduction" (9 March), at http://plato.stanford.edu/entries/abduction/, accessed 22 January 2014.

Ian Steadman, 2013. "Big data and the death of the theorist," *wired.co.uk* (25 January), at http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory, accessed 22 January 2014.

Anselm L. Strauss and Juliet Corbin, 1990. *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, Calif.: Sage.

Nassim Taleb, 2010. *The black swan: The impact of the highly improbable*. New York: Random House.

Barbara F. Walter, 2013. "Syria The four things we know about how civil wars end (and what this tells us about Syria)," *Political Violence @ a Glance* (18 October), at http://politicalviolenceataglance.org/2013/10/18/the-four-things-we-

know-about-how-civil-wars-end-and-what-this-tells-us-about-syria/, accessed 22 January 2014.

Wikipedia, 2013. "PRISM (surveillance program)," at http://en.wikipedia.org/wiki/PRISM_(surveillance_program), accessed 22 January 2014.

Editorial history

Received 22 December 2013; revised 21 January 2014; accepted 22 January 2014.

In the realm of Big Data ...
by Antony Bryant and Uzma Raja.
First Monday, Volume 19, Number 2 - 3 February 2014
http://journals.uic.edu/ojs/index.php/fm/article/view/4991/3822
doi: http://dx.doi.org/10.5210/fm.v19i2.4991.